Legal protection of synthetic data for artificial intelligence applications

Artificial Intelligence (AI) is used more and more in many businesses, in research but also in everyday life. AI applications can recommend online purchases to users, they can support doctors in the diagnosis of diseases such as cancer or they can be used to autonomously drive cars. The range of applications is virtually limitless. But almost all types of AI have one thing in common: they rely on vast amounts of data for the training. Training is what enables an AI to fulfill the task. Without adequate training the application cannot distinguish skin cancer from other irregularities on the patient's skin and an autonomous car may not be able to avoid a collision.

In many cases the required data for the training comes from real world scenarios. It is called 'original data'. Original data is collected based on user activities on the internet, derives from pictures of skin cancer or is based on sensor data from cars driving through our cities. But quite often this data is insufficient because there are not enough data sources or the data is incomplete because information is missing. These insufficiencies can be overcome by the use of synthetic data. Synthetic data is created by specialized AI applications based on data from real life scenarios. It can be used to provide a broader training basis for medical applications and autonomous driving. It can also be used to anonymize sensitive information. In the process of creating synthetic data, called 'synthetization', information that may identify individuals can be removed in a way that keeps valuable information intact for the training but protects the personal rights of individuals. This can be particularly valuable in medical fields.

Because of the opportunities that synthetic data is offering, it is valuable. It also has the potential to allow companies without sufficient resources for the collection of original data, to obtain training data for new and innovative applications. But even though there is value present in the synthetic data, the law does not expressly say who owns the data. And without clear rules on ownership, companies can have an incentive not to share the data with other possible users which will reduce the availability of possible applications. There are laws in the EU and in the EU Member States that apply to training data for AI. But those rules apply to data that comes directly from humans (e.g. data protection laws, copyright) or that are collected from sensors (e.g. the new EU Data Act). Since the technological development has been moving fast in the past years, the necessity for protection of synthetic data has been overlooked. But since there are estimates that in 2024 up to 60% of all training data will be synthetic data, a solution to this problem is required.

This project will analyze the current legal framework for the protection of data, conduct an empirical study with companies in Poland and Germany to assess the right protection of synthetic data and will propose a new or improved legal framework. The new framework is meant to grant ownership in synthetic data and in this way provide companies with the means and the incentives to create and use such data for new applications in a wide array of fields of technology.